**5. Pendekatan Genre dan Domain Spesifik**

Pada bagain ini, penulis membahas berbagai riset tentang peringkasan dalam berbagai genre dan domain yang berbeda, yaitu: domain medis, serta genre artikel jurnal, email, blog, dan percakapan.

**5.1 Peringkas Medis**

Peringkasan (*Summarization*) dalam bidang medis merupakan contoh yang patut diperhatikan dari sebuah aplikasi dimana pendekatan peringkasan yang umum saja tidak berlaku. Algoritma peringkasan dikembangkan dengan maksud penggunaan yang ditentukan seperti untuk membantu doctor dalam memutuskan pengobatan, mengulas penelitian terbaru yang relevan pada pasien tertentu atau untuk membantu pasien dan keluarganya mengakses informasi yang bersakutan pada kondisi atau penyakit.

Dalam komunitas medis, jumlah jurnal dan artikel yang relevan sangatlah besar yang mana membuat para dokter maupun pasien dan keluarga pasien kesulitan dalam mencari informasi yang mereka butuhkan. Disinilah peringkasan digunakan untuk membantu pasien dalam menentukan artikel mana yang pas dari hasil pencariannya di internet, begitu juga dengan para dokter dengan jurnalnya dimana hal ini membantu mereka untuk *browsing* lebih efisien.

Sebuah ringkasan bisa disesuaikan pada tipe user, baik pasien maupun penyedia layanan kesehatan. Konten ringkasan bisa disesuaikan lebih mendalam dengan record pasien, seperti diagnose dan beberapa hasil tes yang mana bisa memengaruhi apa yang relevan untuk dokter dari pasien maupun pasien dan keluarganya itu sendiri. End user (konsumen dan penyedia layanan kesehatan) dalam system peringkas mungkin mencari informasi tentang masalah spesifik ataupun browsing indormasi yang terkait tentang masalah pasien. Dua tugas berbeda ini mungkin juga memengaruhi jenis ringkasan yang disajikan.

Konsep medis dalam Unified Medical Language System (UMLS) bisa secara otomatis menghubungkan term dari inputan, semisal untuk inputan penyakit Hodgkin juga mendaftarkan Hodgkin’s sarcoma, Hodgkin lymphoma dan Lymphogranulomatosis, dimana semua term bisa digunakan untuk mengarahkan pada penyakit itu sendiri.

**5.1.1 Peringkasan Dokumen Layanan Kesehatan untuk para Pasien**

Para pasien berserta keluarganya mencari informasi dengan cara yang agak berbeda dengan pemberi layanan. Seringkali query yang mereka masukan tidak tepat dengan informasi spesifik yang mereka inginkan. Masalah ini tidak hanya terbatas pada pencari informasi medis, tetapi para peneliti juga.

Centrifuser adalah peringkas yang bertujuan membantu konsumen dalam pencarian informasi mereka dengan membuat tiga komponen ringkasan berbeda yang membentuk gambaran dan secara khusus menargetkan kebutuhan kognitif konsumen. Tiga komponen itu terdiri dari (1) hyperlink ke topik yang terkait query untuk memudahkan dalam navigasi dan perumusan query, (2) gambaran umum tingkat tinggi tentang kesamaan dalam dokumen, dan (3) deskripsi perbedaan antara dokumen yang diambil untuk membantu pencari memilih item yang relevan.

Ringkasan Centrifuser bisa digolongkan sebagai multi dokumen dan query-focused dimana peringkas memilih informasi relevan pada banyak dokumen menggunakan query untuk mengakses segmen topik dalam dokumen yang kemungkinan besar berisi informasi yang terkait. Centrifuser juga membuat persamaan dan perbedaan antara dokumen yang dia sedang ringkas.

**5.1.2 Medical Journal Article Summarization**

TAS sistem peringkas jurnal, dikembangkan sebagai bagian dari PERSIVAL, perpustakaan sepsifik digital seorang pasien. TAS menghasilkan sebuah ringkasan untuk sekumpulan dokumen yang terambil sebagai relevan pada query pengguna. Jadi, TAS masuk kedalam golongan peringkasan berfokus query dan multi dokumen. TAS berbeda dari kebanyakan peringkas dimana dia tidak menggunakan pengekstrakan kalimat, tapi malah mengekstrak informasi dari unit sub-sentensial untuk mengisi template yang telah ditentukan sebelumnya, menyusunnya, dan menghasilkan kalimat ringkasan. Ringkasan yang dihasilkan oleh TAS adalah briefing yang berisi hasil yang dilaporkan dalam studi klinis. TAS menggunakan informasi yang diekstrak dari record pasien untuk mencari saringan dari artikel, jadi membuat ringkasan hasil artikel jurnal yang relevan untuk dokter yang disesuaikan pada pasien yang ditanganinya. Record pasien yang digunakan TAS, tersedia untuk PERSIVAL dari sistem record pasien elektronik WebCIS (Web-based Clinical Information System).

Untuk membuat ringkasan. TAS pertama secara otomatis menggolongkan artikel berdasar maksud utama klinisnya sendiri. Tipe artikel khusus yaitu diagnosa, prognosis, dan perawatan. Dalam tahap pemilihan konten, sekumpulan template, masing-masing mewakili sebuah parameter (e.g., dada sakit), relasi (e.g., asosiasi), dan pencarian (e.g., angina tidak stabil), diinstansiasi pada setiap artikel input. Pada tahap ini, hasil yang relevan diekstraksi memanfaatkan struktur artikel medis, yang masing-masing selalu memiliki bagian hasil. Template yang tidak spesifik untuk pasien dilepaskan dengan mencocokkan parameter terhadap informasi yang ditemukan dalam catatan pasien. Misalnya, template yang mewakili hasil "Sakit di dada dikaitkan dengan angina tidak stabil" hanya akan disertakan dalam ringkasan untuk pasien dengan nyeri dada. Selama tahap organisasi konten, template yang relevan dikelompokkan ke dalam unit yang berhubungan secara semantis, dan diurutkan. Akhirnya, TAS menggunakan generator bahasa berbasis frasa untuk menghasilkan ringkasan yang baik.

Yang dkk. juga menggambarkan sistem peringkas multi-dokumen query-focused. Mereka menggunakan ekstraksi kalimat dengan gaya yang sama seperti peringkas umumnya yang dijelaskan sebelumnya, namun sistem mereka sangat berbeda dalam upayanya meringkas semua informasi yang berkaitan pada gen, menspesifikasikan pada query pengguna. Pencarian untuk informasi yang relevan dilakukan di semua PubMed, gudang penyimpanan data abstrak medis secara komprehensif. Dengan permintaan yang mengandung gen yang menarik, sistem mereka mengembalikan daftar peringkat semua kalimat di semua artikel yang relevan dan pengguna dapat memilih berapa banyak kalimat yang ingin mereka lihat. Jumlah kalimat untuk setiap gen bervariasi dari satu sampai 30.216. Seperti Elhadad dkk., Yang dan rekan penulisnya juga menggunakan informasi semantik yang tersedia di bidang medis, namun mereka menggunakannya untuk mengekstrak fitur untuk pengelompokan dan untuk mengekstraksi kalimat yang relevan. Pada tahap pertama sistem mereka, mereka membuat database dari semua kalimat yang diindeks oleh gen tikus yang menggunakan gen / protein mereka yang disebut sistem pengenalan entitas. Mereka juga menyimpan bidang seperti judul MESH (yaitu kata kunci) dan tanggal publikasi untuk setiap kalimat. Sistem ekstraksi kalimat mereka menggunakan seperangkat fitur untuk mencetak kalimat untuk dimasukkan dalam ringkasan, termasuk beberapa berdasarkan informasi semantik domain tertentu. Mereka memodelkan ekstraksi kalimat mereka pada paradigma Edmundson dan ini, satu fitur adalah frase petunjuk yang dapat menunjukkan pentingnya (misal "Dalam kesimpulan"). Kumpulan fitur kunci lainnya mencakup lima fitur deskriptif yang mewakili kelompok gen dalam database (misalnya, istilah MESH, kata-kata deskriptif yang menunjukkan di mana gugus tersebut jatuh dalam ontologi gen). Cluster gen dihitung dengan menggunakan algoritma clustering standar pada vektor gen yang terdiri dari judul MESH, istilah ontologi gen yang terkait dengan gen, dan kata-kata dalam kalimat yang terkait dengan setiap gen serta kalimat yang tepat sebelum dan setelah hukuman sasaran. Fitur lain untuk ekstraksi kalimat mencakup jumlah gen yang dijabarkan kalimat, kata kunci domain spesifik, panjang tak rata, dan masa lalu dari artikel di mana kalimat tersebut terjadi. Skor menghasilkan rangkuman kalimat untuk dimasukkan dalam ringkasan dan dibiarkan kepada pengguna untuk menentukan panjang ringkasan.

**5.2 Ringkasan Artikel Jurnal dalam Domain Non-medis**

Artikel Jurnal medis sangatlah penting bagi dokter dan para peneliti. Untuk mencari informasi penting dalam artikel jurnal tersebut, terdapat beberapa pendekatan yang dapat digunakan. Pada bagian ini, mereke mensurvei sebuah pendekatan dengan *rhetorical status* dan penggunaan informasi tentang kutipan.

**5.2.1 Penggunaan *Genre-speciﬁc Rhetorical Status***

Teufel dan Moens mengusulkan skema anotasi untuk *rhetorical status* dengan kategori AIM, OWN, CONTRAST, dan BASIS. Orang-orang dapat menggunakan skema anotasi tersebut untuk menganotasikan setiap kalimat pada artikel dengan salah satu kategori *rhetorical status*. Awal dari penggunaan rheteorical status, yakni dengan mencantumkan kalimat yang memenuhi AIM, CONTRAST, BASIS dan BACKGROUND untuk setiap *paper* yang memberikan ringkasan generik yang baik dari sebuah artikel ilmiah. Biasanya, ada sedikit kalimat AIM, CONTRAST dan BASIS di *paper* dan semuanya biasanya terdapat di dalam ringkasan. Kalimat BACKGROUND, biasanya ditampilkan lebih banyak sehingga mungkin menjadi masalah ketika membatasi panjang ringkasan. Untuk mengatasi masalah tersebut, diperlukan satu klasifikai lagi yang dapat membedakan kalimat ringkasan dengan kalimat lainnya. Klasifikasi ini hanya digunakan untuk memilih kategori yang paling tepat dan kata kunci yang digunakan di seluruh kinerja *summarizer*.

**5.2.2 Exploitasi Link Kutipan Antar Paper**

Menspesifikan genre pada kesimpulan artikel dapat memberikan nilai dan sumber informasi yang tidak tersedia pada genre lain. Terdapat tiga tipe *extractive summaries* berdasarkan analisis link kutipan yang telah diusulkan, yaitu *overview research area*, *impact summary*, dan *citation summary*.

Menulis overview dari beberapa artikel merupakan hal yang sulit bagi orang-orang. Nanba dan Okumura mengusulkan system untuk memfasilitasi menulis overview, dimana juga menggambarkan hubungan antar paper dan menyediakan informasi tentang bagaimana paper tersebut berhubungan. Mereka mengembangkan sebuah peraturan pada sistem untuk mengidentifikasi sumber area pada paper.

*Impact summary* di definisikan oleh Mei dan Zhai sebagai *task* untuk mengekstrak kalimat dari paper yang mewakili konten yang paling berpengaruh dari paper tersebut. Mereka menerapkan pendekatan model Bahasa untuk menyelesaikan *task*. Model bahasa ini memberikan cara untuk menilai kalimat penting pada artikel. Kalimat penting adalah kalimat yang berisikan informasi serupa dengan paper yang dibahasa saat merujuk ke paper asli. Untuk menjelaskan pentingnya setiap kalimat dalam artikel yang dirangkum, mereka menggunakan probabilitas kata dari artikel dari artikel tersebut. Skor akhir sebuah kalimat adalah kombinasi linear dari *impact importance* yang berasal dari Kullback-Leibler *divergence* dan intisari yang penting dari probabilitas kata yang diinput dalam artikel.

*Citation summarization* adalah pendekatan untuk meringkas sebuah artikel, berdasarkan bagaimana paper lain merujuk pada artikel tersebut. *Citation summarization* tidak menggunakan seluruh teks pada artikel. Qazvinian dan Radev mengusulkan untuk meringkas *reference area* dalam artikel lain yang berhubungan dengan target paper yang akan diringkas. Terdapat pengulangan input yang banyak karena banyaknya paper yang merunjuk pada paper utama, seperti pendekatan, hasil, kontribusi utama, dsb. Melihat karakteristik data yang seperti itu, digunakannya metode *clustering* untuk meringkas paper tersebut. Untuk menemukan representasi kalimat setiap cluster digunakannya metode berbasis grafik, kemudian untuk mengevaluasinya digunakan metode manual *Pyramid evaluation. Citation summaries* lebih susah untuk dibaca dibandingkan dengan *impact summaries* karena *Citation summaries* mencampur informasi tentang bagaimana paper tersebut diringkas, dimana situs artikel tersebut dijelaskan prosesnya dengan sendirinya pada artikel.

Seluruh pendekatan tersebut menunjukkan banyaknya hal yang ada pada peringkasan jurnal. Pendekatan yang sangat berbeda dari generic summarization telah diusulkan untuk mengeksploit struktur dan spesifik karakteristik dari artikel jurnal dan hal ini sungguh menarik. Diperlukannya proses lanjutan yang menggunakan pendekatan baru yang lebih baik.

5.3 Email

Penelitian dalam peringkas email mempunyai rentang dari peringkas email tunggal sampai ke peringkas koleksi email seperti kotak email atau thread email/ percakapan email. Untuk peringkas email tunggal cukup berinti pada topik pesan. Hal ini mempunyai kegunaan yaitu untuk memprioritaskan email masuk, menentukan kapan diperlukan balasan segera dan mencari pesan lama. Untuk kumpulan email (kotak email), peringkas dapat menyediakan sebuah penjelajah untuk membantu user menemukan beberapa email yang diinginkan. Untuk thread email, peringkas digunakan untuk menyampaikan intisari pokok dari percakapan email. Pada kasus ini, peringkas harus sensitif terhadap karakter unik dari email, atau jenis linguistik yang berbeda yang menunjukkan karakteristik teks tulis maupun percakapan lisan.   
**5.3.1 Peringkas dari Pesan Email Tunggal**

Peringkas dari pesan email tunggal dapat dicapai dengan memilih frase nomina yang menunjukkan topik dari email tersebut. Pendekatan ini dilakukan dalam penelitian Gister [141, 208], sebuah sistem yang menggunakan kombinasi filtering linguistik dan machine learning untuk memilih frase nomina untuk summary. Pr3sh1r4m

Peneliti di Microsoft Research [42] mengembangkan sebuah sistem menggunakan metode machine learning Support Vector Machines untuk klasifikasi kalimat dalam email sebagai perintah kerja atau bukan. Hasil terbaik mereka dicapai dengan semua fitur: fitur pesan-specific (seperti, jumlah alamat), fitur superficial (seperti, n-grams) dan fitur linguistik (seperti, fitur struktural yang dapat diturunkan dari tagging atau parsing part-of-speech). Namun perbedaan hasil dengan menggunakan fitur deep linguistik dan surface linguistik sangat kecil, sehingga menunjukkan bahwa sistem mereka akan tetap mendapatkan hasil bagus meski tanpa parsing.  
 Peringkas full dokumen juga telah digunakan untuk peringkas pesan email. Lam et al. [99] menggunakan sistem peringkas IBM [17] untuk peringkas pesan email. Akan tetapi, mereka menggunakan pre-prosesing yang menghapus header tertentu, teks kutipan, informasi terusan, dan tanda tangan email. Mereka menggunakan proses yang mirip dengan peringkas pada email thread.

**5.3.2 Peringkas dari Email Thread**

Pada awalnya, Nenkova dan Bagga [151] mengembangkan sebuah sistem untuk membangkitkan precakapan email dalam diskusi yang tersimpan. Mereka menggunakan extractive summarizer untuk ringkasan 2 level pertama dari pohon percakapan email, sehingga menghasilkan ringkasan ikhtisar yang relatif pendek. Untuk pesan root, mereka mengekstraksi pesan terpendek dengan jumlah nomina terbanyak. Untuk pesan berikutnya, mereka mengekstraksi kalimat dengan kata yang paling cocok dengan pesan root. Mereka menggunakan corpus percakapan email yang cukup pendek, jadi pendekatan ini bekerja dengan baik.

Rambow et al [177] juga menggunakan extractive summarization. Akan tetapi, mereka berfokus pada sifat dialogis dari email. Mereka bereksperimen dengan 2 extractive summarization berbasis machine learning. Eksperimen pertama menggunakan fitur yang mirip dengan peringkas teks lain dan menganggap percakapan masukan sebagai dokumen teks biasa. Peringkas kedua bergantung pada fitur email spesifik sebagai tambahan dari fitur tradisional, termasuk fitur terkait percakapan dan struktur email (seperti jumlah pembalas email, kesamaan kalimat dengan subjek, dll). Menggunakan Ripper, pengklasifikasian rule-based, hasil mereka menunjukkan bahwa fitur dengan full set menghasilkan peringkas terbaik.

Penelitian [140, 187] mengidentifikasi pasangan pertanyaan-jawaban dalam percakapan email. Sedangkan penelitian [132] meneliti integrasi pasangan tersebut dalam peringkas ekstraksi dalam email. McKeown et al. meneliti dengan berbagai metode untuk mengintegrasikan pasangan tersebut dalam peringkas email, dan menunjukkan bahwa semua metode tersebut menghasilkan peningkatan yang signifikan terhadap pendekatan ekstraktif saja.

Peringkas percakapan email juga dapat digunakan untuk membantu user dalam mengejar ketinggalan pesan dalam diskusi panjang, dan kemudian membuat respon. Penelitian Wan dan McKeown [215] menunjukkan bahwa pendekatan berbasis centroid mempunyai hasil yang baik dalam mengangani hal ini. Intuisi metode ini mirip dengan yang dibahas di 5.3.1.

Dari penelitian diatas dapat dilihat bahwa secara keseluruhan, memperhatikan karakteristik yang spesifik untuk email akan menghasilkan sistem peringkas yang lebih baik. Secara umum, machine learning menunjukkan bahwa menggunakan fitur linguistik ditambah dengan fitur spesifik email menghasilkan hasil terbaik.

**5.3.3 Peringkas dari Arsip Email**

Penelitian yang penulis sampaikan sejauh ini telah membangkitkan ringkasan untuk berkas individu. Namun, arsip dan kotak email terdiri dari banyak berkas dan user dapat kesulitan untuk mengatur kotak email mereka sehingga mudah mencari email atau berkas yang diinginkan. Newman dan Blitzer [156] menawarkan sebuah sistem yang dapat digunakan untuk menjelajah kotak email yang terbentuk dari teknik peringkas multi-dokumen. Sistem mereka mengandalkan fitur seperti: fitur spesifik email dan kesamaan dari kalimat dengan cluster centroid, untuk membangkitkan ikhtisar maupun ringkasan.

Carenini et al mengaplikasikan pendekatan yang lebih terkini menggunakan analisis berbasis graf untuk kutipan dalam email dengan menghitung graf dari bagaimana setiap email tunggal menyinggung langsung email lain. “Kata petunjuk” ditetapkan sebagai stem dari kata dalam email yang diulang-ulang dalam node parent atau child dari graf kutipan. CWS (Clue Word Summarizer) menghitung skor kalimat untuk dimasukkan berdasarkan frekuensi kata-kata petunjuk yang dikandungnya. Dibandingkan dengan penelitian MEAD [56], dan versi modifikasi dari ekstraktor kalimat berdasarkan statistik oleh Rambow et al. [177], serta kombinasi dari CWS dengan MEAD menunjukkan bahwa peringkas kata petunjuk bekerja sama baiknya atau lebih baik dari pendekatan sebelumnya.

**5.4 Ringkasan Web**

Banyak informasi dalam web memotivasi untuk melakukan penelitian dalam meringkas informasi di dalamnya. Sistem yang dikembangkan untuk meringkas web memperhitungkan karakteristik dari web dan data yang dihasilkan mempunya kualitas yang bagus serta ringkasan yang akurat. Sekarang ini kebanyakan penilitian berfokus ke ringkasan halaman web seperti blog.

**5.4.1 Ringkasan Halaman Web**

Ringkasan dalam halaman web dipengaruhi konten atau konteks web. Penelitian ini menggunakan sumber daya yang besar seperti salah satunya DMOZ (dmoz.org). Awalnya meringkas web hanya berdasarkan kontennya saja. Untuk membuat sebuah ringkasan dari sebuah konten, Berger dan Mittal menggunakan mesin terjemah statistic. Untuk menghasilkan ringkasan, mereka menggunakan dua model, yang pertama menentukan kata-kata ringkasan dan yang kedua menentukan urutan. Buyukkokten dkk. fokus pada pembuatan ringkasan panjang yang dapat digunakan dalam genggaman. Mereka menggunakan metode peringkas yang relatif sederhana, memilih kalimat yang paling penting dari segmen wacana dalam halaman web.

Penelitian selanjutnya ialah membuat ringkasan dari sebuah konten dan konteks web. Untuk membuat rangkuman, sistem tersebut menggunakan pencarian url, memilih semua kalimat yang berisi link ke URL dan kalimat terbaik itu diidentifikasi menggunakan heuristik. Delort dkk. menggunakan prosedur yang sangat mirip untuk memilih kalimat konteks. Mereka memperpanjang pendekatan yang lebih tua melalui algoritma yang memungkinkan seleksi Dari kalimat yang mencakup sebanyak mungkin aspek halaman web dan itu adalah topik yang sama.

Sun et al. [197] menawarkan konteks yang lain: penggunaan klik per data. Mereka menggunakan data dari mesin pencari Microsoft untuk mengumpulkan tiga komponen (u, q, p) mewakili kueri (q) keluaran oleh pengguna (u) siapa yang mengklik ke halaman (p). Dalam karya ini, ringkasan dibuat dengan memilih kalimat dari halaman web. Mereka bereksperimen dengan dua metode, salah satunya menggunakan adaptasi metode Luhn [111], menghitung faktor signifikansi kata dengan campuran tertimbang TF \* IDF dihitung berdasarkan frekuensi pada halaman web dan TF IDF dihitung berdasarkan frekuensi dalam kumpulan kueri

**5.4.2 Ringkasan situs online interaktif**

Metode yang digunakan oleh Zhou and Hovy untuk ringkasan ini mirip dengan metode meringkas email dan pembicaraan. Penelitian mereka berfokus pada identifikasi kedekatan yang berpasangan guna mengidentifikasi siapa bicara ke siapa dan topik apa yang sedang dibacarakan. Untuk ringkasan dalam blog ada beberapa metode seperti, menghitung frekuensi kata untuk menentukan kalimat yang penting.

**5.5 Peringkas ucapan**

Beberapa teknik yang dikembangkan dalam meringkas teks berhasil diterapkan pada peringkas ucapan tanpa banyak modifikasi. Teknik yang banyak digunakan untuk meringkas transkrip adalah Maximal Marginal Relevance dan Latent Semantic Analysis. Variasi fitur frekuensi (frekuensi total, TF \* IDF) dan fitur posisi, serta fitur panjang, serupa dengan yang dikembangkan untuk peringkas teks.

Terdapat beberapa aspek dari ringkasan ucapan yang menyebabkan kebutuhan penelitian dalam arah yang belum dijelajahi dalam penyusunan teks. Memerlukan verifikasi dan analisis lebih lanjut yang menunjukkan bahwa kinerja rangkuman yang sangat baik dapat dicapai berdasarkan fitur akustik saja, tanpa adanya transkrip atau fitur teks.

**5.5.1 hubungan dengan kesalahan dari pengenalan ucapan otomatis**

Dalam situasi yang paling realistik, transkrip manual dari input audio yang diringkas tidak tersedia. Dalam hal ini, Automatic Speech Recongnition (ASR) digunakan untuk mendapatkan transkrip teks. Bergantung pada jenis ucapan, ASR bisa salah, dengan tingkat kesalahan bisa 10% untuk ucapan baca yang lebih formal hingga 40% untuk percakapan spontan. Sistem ringkasan ucapan dimaksudkan untuk menghasilkan output tekstual perlu mengembangkan teknik yang tidak informatif namun juga memiliki tingkat kesalahan pengenalan lebih rendah.

Untuk memberi gambaran bagaimana penggunaan transkrip otomatis mengubah kejernihan dan isi ringkasan, terdapat contoh Murray :

Human : Percobaan terdiri dari memimpin subjek untuk percaya bahwa dia sedang berbicara dengan komputer, kemudian mengalami "komputer" rusak dan diganti dengan manusia.

LSA; manual transcript : Seharusnya saya mengatakan bahwa sistem rusak dan ini adalah tiga tugas yang tersisa yang akan dia selesaikan dengan manusia.

LSA; automatic transcript : Seharusnya dibalik sistem yang diruntuhkan dan kemudian ini akan tersisa tiga tugas yang akan dia selesaikan dengan manusia.

Sebagai alternatif, upaya yang dapat difokuskan untuk memperbaiki kinerja ASR sebelum dilakukan ringkasan. Mengubah modalitas output juga dapat menghapus kebutuhan untuk secara eksplisit menangani kesalahan ASR. Misalnya, kesalahan tidak akan menjadi masalah jika rangkuman akhir diberikan dengan memutar cuplikan audio bertekad untuk menjadi penting oleh algoritma yang kuat yang tidak sensitif terhadap kesalahan transkripsi.

**5.5.2 Hubungan dengan ketidaklancaran : mengisi jeda dan pengulangan**

Ucapan spontan ditandai dengan adanya jeda yang penuh ("um", "eh", "well") dan pengulangan. Presentase kata – kata yang tidak disengaja dalam percakapan formal bisa tinggi: 15-25% dari total kata yang diucapkan. Pertimbangkan contoh ujaran percakapan awal Zechner dan versi pembersihannya:

A: well I um I think we should discuss this you know with her

A’: I think we should discuss this with her

Dalam karya Zhu dan Penn mengemukakan sebuah saran untuk berbagai perlakuan ketidaklancaran. Mereka berpendapat bahwa alih-alih menyingkirkan ketidaklancaran dari input, sistem harus menggunakan kehadiran mereka sebagai fitur dalam mesin model pembelajaran untuk ringkasan ekstraktif. Dalam eksperimen mereka, menggunakan fitur ketidaklancaran - jumlah pengulangan dan jeda yang terisi - menyebabkan peningkatan kecil kurang dari 1% skor ROUGE.

**5.5.3 Unit Untuk Ekstrasi Konten**

Beberapa pendekatan telah dikembangkan untuk segmentasi otomatis ucapan (transkrip) ke dalam kalimat. Secara bersamaan mendeteksi batas-batas kalimat dan disfluensi, memungkinkan transformasi percakapan percakapan asli yang sulit dibaca menjadi bentuk teks yang lebih biasa:

Original : but uh i i i i think that you know i mean we always uh i mean ive ive

had a a lot of good experiences with uh with many many people especially

where theyve had uh extended family and i and an- i i kind of see that that

you know perhaps you know we may need to like get close to the family

environment.

Processed : But ive had a lot of good experiences with many people especially where

theyve had extended family. I kind of see that perhaps we may need to get

close to the family environment